iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** III **Month of publication:** March 2026

DOI: https://doi.org/10.22214/ijraset.2026.78256

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

Literature Survey on Comparative Analysis of Machine Learning Techniques

Om Deokar[1], Rajeev Patil[2], Vijaykumar P. Mantri[3]
Department of Computer Engineering MIT Academy Of Engineering, Pune, India

Abstract: ***In the educational field predicting student performance importance and can provide results that can assist educators to plan accordingly. When students who are at risk of not performing well are identified beforehand strategies can be planned accordingly. For this purpose student performance prediction is very useful however, manual assessment of previous grades is a tedious and time consuming task. We have presented a detailed study where ML techniques are implemented to predict the student academic performance. Many supervised machine learning algorithms that are tree based, probabilistic, distance based and margin based and many others that are implemented are studied. Techniques like feature selection and preprocessing reduce noise and bias. The paper also explains how accurate performance prediction can help educational institutions in planning academics and guiding students that need support.***

Index Terms: Student Academic Performance Prediction, Machine Learning, Data Mining, Algorithms, Predictive Analytics, Academic Performance Analysis

I. INTRODUCTION

Data related to education and digital learning platforms is available in abundance .Academic records are present in many forms. It is an inefficient and tedious job to manually analyze this data and may cause errors too. A solution to this problem can be ML techniques as they utilize data mining for the purpose of prediction. Accurate prediction is essential for educators to identify the students who are probable to underperform and design a plan to implement corrective measures. Emotional and psychological factors also influence a stu-dent's performance like socio-economic background, learning habits and past performance which too needs to be analyzed. Machine Learning algorithms are designed in a manner such that they handle these relationships too. Predictive models are trained on historical data, meaningful patterns are extracted that help in decision-making process.ML techniques help to improve accuracy. Digital transformation has changed the education field and has led to generation of tremendous amount of data.ML techniques are capable of handling this large amount of data and the variations present within this data. Models are trained on historical student data, meaningful patterns are discovered that support to make decisions. The integration of ML techniques also helps to improve the pre-diction accuracy, reduce bias, and design data driven educational strategies. In recent years, there has been a rapid digital transformation in the educational field due to the adoption of learning platforms in online mode,LMS and academic sys-tems. These platforms continuously generate large volumes of structured and unstructured data related to students' academic performance, attendance, learning behavior, assessment results and interaction patterns. The extraction of meaningful insights from this data is a challenge for institutions that aim to boost learning outcomes reflected through student success rates.

The traditional methods may rely only on exam scores and such methods do provide a basic understanding of student's academic status yet they may not capture the complex fac-tors that influence their academic performance. Factors like attendance consistency, learning habits and socio economic background play a significant role which may be missed.

Moreover, student performance prediction is important to improve the learning effectiveness and overall education qual-ity in educational institutions. The effectiveness of teaching learning can be improved when gaps and identified and remedies are designed to them on time. Educational institutions are now using intelligent decision support systems that help faculties to monitor student performance at an individual level and design customized teaching strategies for them. Scalability is required in institutions where analysis of data of thousands of students is simultaneously done.Machine learning based systems learn from these patterns and can handle diverse curriculum structures across different institutions.

Fig. 1 illustrates the of machine learning techniques considered in various studies that are organized into major cat-egories. There exist four machine learning techniques namely supervised learning, unsupervised learning and deep learning and dimensionality reduction methods. Supervised learning methods depend on labeled data to predict outcomes they include Na¨ıve Bayes (NB),Random Forest (RF),k-Nearest Neighbors (KNN) and Support Vector Machines (SVM).

©IJRASET: All Rights are Reserved

Unsupervised learning methods are used to find hidden

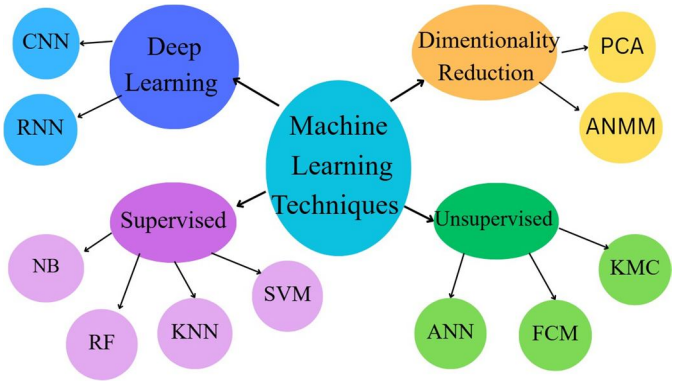


Fig. 1. Machine Learning Techniques

structures in unlabeled data they include Artificial Neural Networks (ANN), Fuzzy C-Means (FCM) and K-Means Clus-tering (KMC).

Deep learning is a subset of machine learning which in-cludes Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) which are useful to recognize pat-terns.

Principal Component Analysis (PCA) and Adaptive Non-linear Mapping Methods (ANMM) are also implemented in various studies. This classification gives a complete overview of the algorithms implemented and studied.

II. LITERATURE REVIEW

Student performance prediction has become very important in recent times due to the growth of digital academic systems and learning platforms [8].Educational data includes academic scores, attendance behavioral patterns and demographic infor-mation [6].Accurate prediction is a difficult job as multiple factors that influence a student's performance are present [10].Traditional methods that are based on manual evaluation and statistical analysis are limited and cannot handle complex datasets [15].These approaches fail to identify students who are at risk [11].Researchers have explored automated predic-tion techniques using data-driven models [13].Machine learn-ing techniques enables efficient analysis of large educational data [1].These techniques improve the prediction accuracy and help in early academic intervention [14].

Classification-based models are commonly preferred for predicting categorical outcomes such as pass or fail [9].De-cision tree algorithms are widely used because of their simple structure and interpretability [7].They help to identify key attributes that influence student performance, such as atten-dance and internal assessments [2]. Their transparent decision-making process makes them suitable for educational applications [6].Probabilistic models like Naive Bayes were studied because of their simplicity and low computational cost [3].

Despite assuming feature independence, these models show competitive results on preprocessed datasets [5]. They are par-ticularly useful for baseline performance prediction [12].Sup-port Vector Machine models were introduced to handle non-linear patterns in student data [6].Research showed that the performance of SVM performs is good in feature spaces of high dimension [3].Proper kernel selection also improves the classification performance [10].Due to increase in data complexity ensemble learning techniques gained importance in educational data mining [1].Random Forest models improve the accuracy as they combine many decision trees together for analysis [2].These models reduce the overfitting [9].Data preprocessing plays a crucial role in student performance prediction [5]. Techniques like normalization, handling miss-ing value and encoding significantly affects model outcomes [4].Poor preprocessing may cause bias in results and may re-duce the prediction reliability [15].Recent studies highlighted the role of predictive analytics to reduce the dropout rates and improve academic planning [6].Early identification of weak students helps institutions to provide academic support [7]. Such approaches improve overall outcomes of the teaching learning process [11].

Large scale of academic and behavioral data is generated due to rapid digitalization of educational institutions which makes student performance prediction a research problem to work on [8].Datasets that are diverse and massive in nature are generated through modern academic systems that contain data related to students and their academic activities which is very precious to predict future outcomes [6].However,owing to the multidimensional nature of educational data, it is difficult to accurately predict student performance [10].

Early studies on this topic mainly focused on traditon-ally used statistical techniques and manual evaluation which provided limited capability to predict outcome [15].These approaches were unable to capture the relationships among academics,behaviour and other factors [11].Therefore,such tra-ditional methods mostly failed to accurately predict perfor-mance of students and did not provide a clear result to design a strategy [7].

Thus to find better solutions,researchers began to adopt machine learning techniques to perform data-driven student performance prediction [13].

ML models can be trained on educational data collected over a period of time and do not need any extra formulation and they can yet extract patterns from it [1].Several studies have demonstrated that ML techniques are better than statisti-cal methods used traditionally in terms of prediction accuracy [14].

Classification-based machine learning models are widely used for prediction of student academic prediction,as academic outcomes are represented as categorical labels such as pass, fail or grade levels [9].These models provide insights that can be used to design academic planning, counseling and personalized learning strategies [6]. Comparative studies have shown that classification techniques are more effective than regression-based models when discrete performance categories are present [10].

Decision Tree algorithms have been extensively studied as they are interpretable and are able to identify key attributes that affect student performance [7].These models generate hierarchical decision rules that help educators to understand
how factors such as attendance, internal assessments and prior grades influence academic outcomes [2].The transparent struc-ture of decision trees makes them particularly suitable for educational environments where things need to be explained [6]. Probabilistic models such as Naive Bayes have also been applied to student performance prediction because of their simplicity and low computational requirements [3].Despite the assumption of conditional independence among features here,Naive Bayes classifiers have achieved good results on well-preprocessed educational datasets [5].These models are often implemented as baseline classifiers for performance comparison in predictive studies [12].

Support Vector Machines have been introduced to handle decision boundaries that are not linear and feature spaces dimensionally high in educational data [6].Research indicates that SVM classifiers perform effectively when appropriate kernel functions are selected and hyper-parameters are used in an optimized manner [3].Kernel-based learning helps SVMs to model relationships between student attributes and perfor-mance outcomes [10].

Educational datasets are growing in size and complex-ity,ensemble learning techniques have gained significant at-tention in recent studies [1].Random Forest models combine many decision trees for improving the prediction accuracy and reduce variance in the model [2].Empirical evaluations have shown that ensemble-based approaches generally outperform individual classifiers as they improve generalization [9].

Several studies have found the that preprocessing data is important for increasing the effectiveness of prediction models [5].Preprocessing techniques such as normalization, handling missing values and encoding attributes of categorical nature significantly impacts the model performance [4].Inadequate preprocessing may lead to bias and noise which furthur may lead to misleading conclusions [15]. Recent research describes the role of predictive analytics to reduce dropout rates and improving educational outcome [6]. Early identification of low-performing students helps institutions to provide aca-demic support [7]. However, existing literature also points out to challenges like data imbalance,privacy concerns and lack of standardized evaluation frameworks, thus more research is required in this domain [13].

Recent studies have also shown a growing interest in applying advanced ML methods for performance prediction in an more accurate manner in higher education environments [17].There is a huge availability of large and diverse educa-tional datasets for researchers to study different models that are capable of identifying learning patterns that traditional methods may miss out [16].These modern approaches improve both, the prediction accuracy as well as usability in decision-making systems [20].

Several researchers have emphasized that deep learning models can relationships between academic and behavioral features [16].They learn hierarchical representations from student data and show good performance as compared to traditional machine learning algorithms [19]. However, their effectiveness largely depends on data quality,model tuning and availability of sufficient training samples [17]. The integration of predictive analytics into systems is a key application of prediction of student performance research [18].Such systems continuously monitor student progress and thus institutions can intervene before academic difficulties become severe [18]. If students who are at risk are identified early,it improves the retention rate and academic success [17].

Machine learning approaches implemented in hybrid man-ner have overcome the limitations of individual classifiers in educational prediction tasks [19].By combining outputs from multiple models,hybrid frameworks reduce prediction variance across different student populations [16].These approaches have shown to generalize better when applied to datasets that are collected from different academic contexts [19].

Recent literature has also highlighted the importance of explaining concepts in student performance prediction models [20].Explainable ML techniques help educators to understand the reasoning of the predictions made,and make predictions more reliable [20].Transparency is important in educational institutions as ethical considerations and accountability are important here [17].

Overall,contemporary research reflects a shift towards more intelligent,interpretable and scalable prediction frameworks for educational analytics [16].The combination of deep learning, hybrid modeling and explainable AI is a very a promising direction for future student performance prediction systems [18][20].

Several studies have emphasized the importance of se-lecting appropriate machine learning algorithms for student performance prediction [10].Supervised learning techniques like Decision Trees,SVM,Random Forest and Logistic Re-gression are widely adopted due to their strong classifica-tion capabilities [11].Ensemble approaches further enhance prediction stability by combining multiple weak learners to improve generalization performance [12].Feature selection and dimensionality reduction methods are also frequently applied to remove irrelevant attributes and improve computational efficiency [8].Improper model selection or overfitting may reduce the reliability of predictive outcomes and affect institu-tional decision-making [14]. Recent research has increasingly focused on integrating explainable AI techniques to improve transparency in academic prediction systems [1].Interpretabil-ity methods such as SHAP and LIME assist educators in understanding the contribution of different attributes toward performance outcomes [23],[24].The growing availability of large-scale educational datasets and learning management sys-tem logs has enabled more data-driven analysis of student behavior [9].However,challenges such as model bias,fairness concerns and scalability in real-world academic environments remain critical research gaps that require further investigation [28],[34].

Table I shows the different machine learning methods applied on educational data [1]-[15], including supervised models,ensemble techniques and analytical reviews.

TABLE I
SUMMARY OF RELATED WORK (REFERENCES [1]–[15])

Author & Year	Study Focus	ML Techniques Used	Key Findings	Dataset / Paper Link
Ahmed et al. (2025) [1]	Academic performance pre-diction with explainability	Random Forest, XGBoost, Logistic Regression	Interpretability improves de-cision making in institutions	Paper Link https://www.nature.com/articles/ s41598-025-12353-4
Gul et al. (2025) [2]	Comprehensive ML frame-work for prediction	Random Forest, SVM, Logistic Regression	Data-driven framework improves predictive performance	Paper Link https://link.springer.com/article/10.1007/ s10791-025-09585-3
Rahman et al. (2025) [3]	Systematic review of AI in education	SVM, Random Forest, Naive Bayes	Identifies research trends and gaps in performance predic-tion	Paper Link https://scholar.google.com/scholar?q=Artificial+intelligence+in+education+A+systematic+review+of+machine+learning+for+predicting+student+performance+Rahman
Buzducea et al. (2024) [4]	ML for aca-institutional demic decisions	Decision Tree, Random Forest	ML supports institutional planning and academic interventions	Paper Link https://www.mdpi.com/2076-3417/14/6/2412
E. Ahmed (2024) [5]	Student performance predic-tion using ML	SVM, Decision Tree, Random Forest, KNN	Comparative evaluation of ML models for prediction	Paper Link https://www.hindawi.com/journals/complexity/ 2024/5567124/

Munir et al. (2024) [6]	AI integration in digital edu-cation	Random Forest, Logistic Regression	ML techniques enhance dig-ital learning analytics	Paper Link https://www.sciencedirect.com/science/article/ pii/S2666920X24000146
Al Husaini & Shukor (2024) [7]	Factors affecting academic performance	Logistic Regression, De-cision Tree	Behavioral and academic indicators influence performance	Paper Link https://www.researchgate.net/publication/379428942
Yag˘cı (2022) [8]	Educational data mining pre-diction	SVM, Random Forest, KNN, Naive Bayes	High classification perfor-mance achieved using en-semble models	Paper Link https://slejournal.springeropen.com/articles/ 10.1186/s40561-022-00192-z
Rao & Kumar (2021) [9]	Online course performance prediction	Decision Tree, Random Forest	LMS behavioural data im-proves prediction accuracy	Paper Link https://uijrt.com/articles/v2/i11/UIJRTV2I110007.pdf
S¸ekerog˘lu et al. (2021) [10]	Systematic review ML of prediction models	SVM, Random Forest, Logistic Regression	Highlights dataset imbalance and evaluation gaps	Paper Link https://www.mdpi.com/2076-3417/11/16/7376
Albreiki et al. (2021) [11]	Review of ML techniques for prediction	SVM, Random Forest, Naive Bayes	Summarizes ML algorithms used in education prediction	Paper Link https://www.mdpi.com/2227-7102/11/9/552
Namoun & Alshanqiti (2020) [12]	Learning analytics and pre-diction review	Decision Tree, Logistic Regression	Educational analytics improves early detection of performance risks	Paper Link https://www.mdpi.com/2076-3417/10/1/237
Hashim (2020) [13]	Supervised ML performance prediction model	Decision Tree, SVM	Supervised learning models classify academic outcomes effectively	Paper Link https://iopscience.iop.org/article/10.1088/1757-899X/928/3/032019
Enughwure & Ogbise (2020) [14]	ML applications in education review	Decision Tree, Random Forest	ML improves predictive ed-ucational analytics	Paper Link https://www.irjet.net/archives/V7/i5/IRJET-V7I51123.pdf
Altabrawee et al. (2019) [15]	Predicting student academic performance	SVM, Naive Bayes	ML models effectively pre-dict university student out-comes	Paper Link https://journalofbabylon.com/index.php/JUB/ article/view/1515

III. METHODOLOGY

The methodology used includes,collect ion of historical student dataset and applying machine learning algorithms on them for prediction.

Eight Machine learning algorithms namely Linear Regression,Naive Bayes,KNN,Logistic Regression,Decision Tree,Random Forest,SVM and XGBoost were used for machine training.The performance of all models was evaluated using evaluation metrics.

This workflow is used so that the most relevant attributes help to predict outcomes. The used methodology inculcates scalability and adaptability across various academic environ-ments.Additional features like learning styles and beahviour in real time can be included using preprocessing and model training.To minimize overfitting and build a strong model cross-validation strategies are used. This design help proper deployment in real-world academic settings.

Data splitting is implementing using the 50:50 ratio, where 50of the dataset is used for training whereas the other 50prevents bias and improves generalization performance.Grid search is used for the tuning of hyperparameters.

A. Tools and Techniques

The studied systems are implemented using the Python programming language as it extensively supports data analysis and machine learning applications.Widely used libraries in Python like Scikit-learn, Pandas, NumPy and Matplotlib are used to process data, implement model implementation and visualize results.Pandas and NumPy are used for data manip-ulation,handle missing values and numerical computations.For visualization of distributions and performance metrics Mat-plotlib is used.Feature selection methods like correlation anal-ysis are implemented for identification of the most relevant attributes in the dataset.This process improves the model's efficiency and prediction accuracy as irrelavant features are re-moved.Various classification algorithms are trained and tested using labeled student datasets and their performance is sys-tematically compared.Comparative analysis helps to identify the most suitable machine learning models to predict student performance.Figure 2 presents the workflow of the prediction system,starting from students' data collection and preprocess-ing. The dataset is cleaned,normalized and finally training and testing subsets are made from it.

B. Analysis Of System

Data collection/acquisition is the first step where academic records of students are collected from databases or academic information systems of institutions. These records include at-tributes like attendance, internal assessment scores, previous semester grades and demographic information. Once this is collected, the data is preprocessed to maintain quality and consistency. Preprocessing includes data cleaning to remove noise and inconsistencies, normalization for scaling numerical attributes and encoding categorical variables into numerical formats. After preprocessing, the dataset is then passed to machine learning models for training. The trained modelshen according to predefined performance categories such as high,medium or low performance categorize the students.The predicted outcomes are analyzed and interpreted and they help educators understand student's learning patterns and make learning roadmap accordingly.

C. Algorithm Description

Processing the raw dataset to handle missing val-ues,normalize the features and encode categorical at-tributes.Feature extraction and selection techniques are applied for retaining the important attributes are used to predict student performance.The processed dataset is furthur divided into training and testing datasets to implement model eval-uation.Training dataset is then used training the machine learning classifiers which helps the models learn patterns and relationships with data.After training is completed,the models are evaluated using the testing dataset to assess their predictive capability.

Evaluation metrics are calculated for every model and the algorithm which performs best is selected based on these metrics.The finalized model is then implemented to predict student performance for new data in academic institutions.

Algorithm 1 Student Performance Prediction Framework

1: **Input:** Raw student dataset D
2: Handle missing values in D 3: Normalize numerical features 4: Encode categorical attributes
5: Apply feature extraction and selection
6: Split dataset into training set D_{train} and testing set D_{test}
7: **for** each classifier M_i **do** 8: Train M_i using D_{train} 9: Predict using D_{test}
10: Compute Accuracy, Precision, Recall and F1-score
11: **end for**
12: Select best performing model M_{best} based on evaluation metrics
13: Use M_{best} to predict student performance for new data
14: **Output:** Predicted academic performance

Figure 2 presents the workflow of the prediction sys-tem,starting from students' data collection and preprocess-ing.The dataset is cleaned,normalized and then divided into training and testing subsets.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue III Mar 2026- Available at www.ijraset.com

D. Experimental Setup

Real-world student datasets that contain aca-demic, behavioral and demographic attributes are used for experimental analysis. Attributes iike attendance percentage, assignment performance, previous academic results and basic demographic information are present. The dataset is then bifurcated into training and testing datasets using a predefined ratio. Multiple machine learning algorithms are implemented under same experimental conditions for comparison. The system performance is then evaluated using classification metrics like accuracy, precision,recall,F1-score

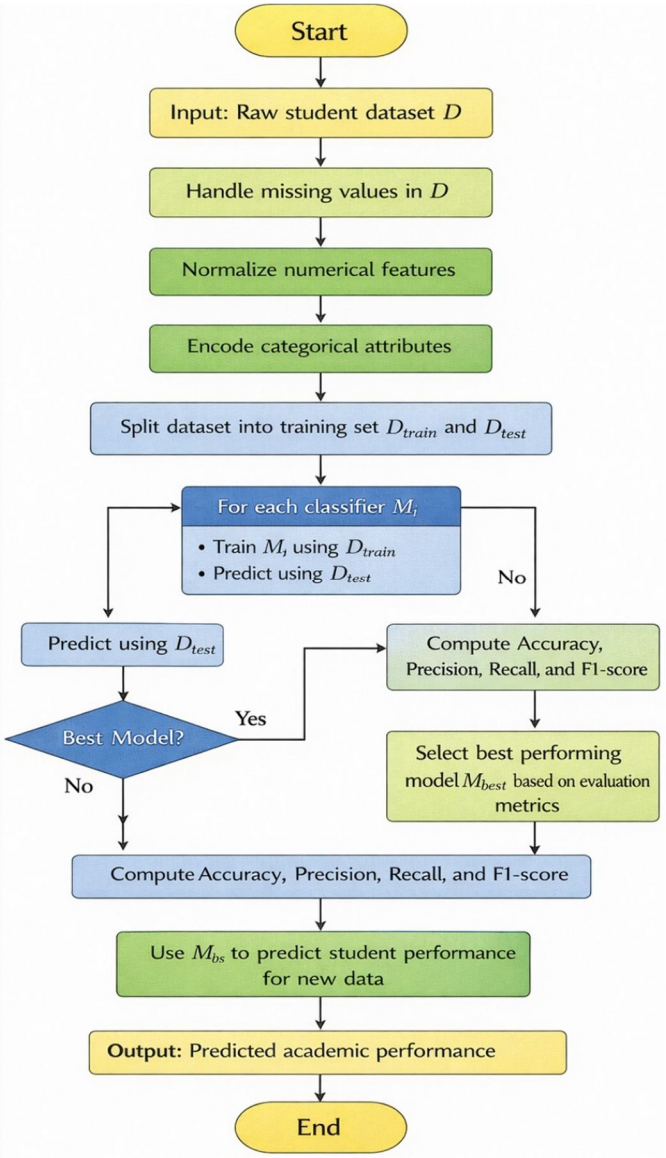


Fig. 2. Algorithm of Model Training and Testing

and confusion matrix analysis. Each model's ability to predict student performance accurately is reflected through these metrics.

E. Evaluation Strategy

The evaluation strategy focuses on assessing both, the overall classification performance and the effectiveness of class-wise prediction Accuracy generally is used as a measure of the model's performance, whereas, precision and recall are used to evaluate the system's capability to correctly identify students who may perform less. Confusion matrix analysis is implemented to analyze misclassification patterns among different performance categories, providing deeper insight into model behavior. To ensure generalization, k-fold cross-validation is applied during model evaluation. This technique validates model consistency across multiple subsets of the dataset and reducesthe dependency on a single data split.Cross-validation also helps to minimize overfitting and improves the reliability of performance comparisons.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

F. System Architecture

Modularity is used in system architecture ,which consists of different stages namely data collection,preprocessing,feature selection, model training, prediction and evaluation.Each stage performs a specific function and interacts with other com-ponents. Data flows sequentially from raw input to final prediction to make system design scalable.Easy integration of additional features,datasets or machine learning algorithms is possible.The complete workflow is represented using system architecture and flow diagrams to explain interaction between different components.It becomes easy to implement and de-ploy structured designs in real-world academic settings.



Fig. 3. Flowchart of Model Training and Testing

Figure 3 presents the sequential data flow while process-ing it for student performance prediction. The process be-gins with gathering student records, demographic information and behavioral data, which undergo data preprocessing. In this stage, missing values are removed or replaced, features are normalized, categorical attributes are encoded and data cleaning is done. Feature selection techniques are applied to get the most relevant features for model training.

After preprocessing and feature selection, the dataset is then divided into training and testing subsets. Various ma-chine learning algorithms including Decision Tree, Random Forest, SVM,k-NN and Naive Bayes are trained using the training data.

The trained models are then tested on the test data and evaluated using metrics like accuracy,F1-score and confusion matrix analysis.

IV. DATASET, PREPROCESSING

The dataset includes student academic and behavioral at-tributes collected from multiple semesters. Preprocessing tech-niques like normalization, handling missing values and cat-egorical encoding are applied. Performance metrics that are used are Accuracy,Precision, Recall and F1-score

For a complete evaluation of prediction performance these metrics are used. An important component of student per-formance prediction is the quality and the structure of the dataset used training and testing the model.The dataset con-sidered in this study consists of academic,behavorial and demographic attributes that are collected over multiple aca-demic years.Features include student attendance percent-age,internal assessment scores,previous semester grades and socio-economic indicators.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

II presents the input attributes and the target variable used for prediction. Features like attendance,academic history assessment performance and study behavior are the factors that affect student academic outcomes.

TABLE II
DESCRIPTION OF STUDENT PERFORMANCE DATASET

Attribute	Description
Attendance	Percentage of classes attended by the student
Internal Marks	Average score obtained in inter-nal assessments
Assignment Score	Score of assignment completion and its quality
Previous Grade	Academic performance in previ-ous semester/examination
Study Hours	Average number of study hours per week
Grade	Final academic performance predicted by the model

Feature engineering plays an important role in transforming raw educational data into inputs for machine learning models. To normalize the features in numerical format and ensure that they contribute uniformly in model training min-max scaling is used. Categorical variables such as gender, department and performance categories are encoded using label encoding techniques. Moreover, derived features like cumulative grade point average (CGPA),improvement trend across semesters and attendance consistency are generated that capture long-term academic behavior. Predictive capability of the model is increased by these derived features.

Data collection is the primary step of the student performance prediction system, the model's accuracy largely depends quality of data that is collected from multiple sources. In this system, data is gathered from multiple sources like academic records, demographic details and behavioral information. Academic data includes marks obtained in previous examinations, attendance records, assignment submissions and internal assessments. Demographic data may include age, gen-der, socioeconomic background and parental education level that may also influence student's learning abilities. Behavioral data includes student engagement like participation in class activities, login frequency to learning platforms and study patterns.

V. MATHEMATICAL FORMULATION

Let the dataset consist of n student records with m at-tributes. Each student record is represented as:

$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\} \quad (1)$$

where $x_i \in R^m$ represents the feature vector of the i^{th}
and y_i represents the corresponding target class.

A. Naive Bayes

Naive Bayes calculates the posterior probability using the Bayes theorem.

$$P(C/X) = \frac{P(X/C)P(C)}{P(X)} \quad (2)$$

The predicted class is

$$\hat{y} = \arg\max_C P(C) \prod_{j=1}^{n} P(x_j/C) \quad (3)$$

B. Linear Regression

Linear regression predicts the output using a linear function

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m \quad (4)$$

here, β represents model parameters.

C. Logistic Regression

Uses the sigmoid function for classification.

$$P(y = 1/x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}} \quad (5)$$

D. k-Nearest Neighbors (k-NN)

Distance between two samples is calculated using Euclidean distance.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \quad (6)$$

Prediction is based on majority voting among the *k* nearest neighbors.

$$\hat{y} = \arg\max_{c} \sum_{i \in N_k} I(y_i = c) \quad (7)$$

E. Support Vector Machine (SVM)

SVM determines a hyperplane that separates classes.

$$f(x) = w \cdot x + b \quad (8)$$

Optimization objective :

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad (9)$$

subject to

$$y_i(w \cdot x_i + b) \geq 1 \quad (10)$$

F. Decision Tree

Decision trees use entropy to measure data impurity.

$$H(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \quad (11)$$

Information gain is calculated as

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \quad (12)$$

G. Random Forest

Random Forest combines predictions from multiple decision trees.

$$\hat{y} = mode\{T_1(x), T_2(x), \ldots, T_K(x)\} \quad (13)$$

H. XGBoost

XGBoost optimizes an objective function and combines loss and regularization.

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (14)$$

where,

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \quad (15)$$

I. Evaluation Metrics

Model performance is evaluated using classification metrics.

Accuracy :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (16)$$

Precision :

$$Precision = \frac{TP}{TP + FP} \quad (17)$$

Recall :

$$Recall = \frac{TP}{TP + FN} \quad (18)$$

F1-score :

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (19)$$

VI. COMPARATIVE ANALYSIS

Comparative evaluation of multiple classification algorithms indicates the effectiveness of ensemble-based approaches, es-pecially Random Forest models.Such models possess a ca-pability to combine multiple decision trees which reduces overfitting. Interpretable models like Decision Trees are also worth using as they identify academic indicators like atten-dance,internal assessments and prior academic performance. These insights are supportive to educators to understand what affects student success and failure.

The results of this work show how the machine learning sys-tems can function as a decision-supporting tool in educational field.When early prediction of student performance is done it helps to intervene and provide personalized mentoring or any other remedy as per requirement.Identifying the students who may be at a risk of performing low early,institutions and educators can take precautionary measures on time to improve the learning outcomes and reducing the dropout rates.This process improves the educational quality.The approaches that are studied also contribute to academic planning and also include intelligent educational analytics.

Inspite of these promising results,the current study does have certain limitations that must be analyzed. Performanceof machine learning models depends on the quality,size and diversity of the dataset that is used for the training and testing purpose.The dataset considered for this study has a limited scope and may not fully represent all educational environments or student.In addition to that,the study focuses mainly on traditional machine learning algorithms and does not include deep learning components that may capture more complex behavioral patterns.Another important limitation is,the limited inclusion of psychological,motivational and emotional fac- tors,which influence student performance but are difficult to collect and measure . Additionally,the integration of behavioral and psychological indicators such as motivation,stress levels and learning prefer-ences can improve prediction effectiveness.Although collect- ing such data is practically challenging,combining quantitative academic data with qualitative behavioral insights can lead to a more holistic understanding of academic performance of students.Advancement in educational data collection andanalytics may help to address these challenges in future studies. Real-time prediction systems when integrated with LMS can improve the practical application of student performance prediction models as they will continuously

track student activities and academic progress throughout the semester.Such systems can provide dynamic risk assessment,which is more beneficial for educators to respond promptly to changes in student behavior and performance.By using real-time analyt-ics, institutions can move from reactive academic support to proactive and preventive intervention strategies.

Over a period of time,such adaptive learning systems can improve student involvement, motivation and long-term academic success. Beyond traditional academic records,data from online learning platforms,discussion forums, assignment submissions and assessment response times can give detailed insights about student learning behavior.Such multimodal data will help models to detect subtle indicators of learning diffi-culties that are not visible through grades alone.

Table III explains the comparative performance of the implemented machine learning models in terms of Accu-racy,Precision,Recall and F1 Score (in %).The results show differences among the algorithms.Linear Regression achieved the lowest overall performance,in Recall (80.31%) and F1 Score (80.21%), which indicates its limitation in classifica-tion tasks. Naive Bayes and KNN showed higher predictive capability,with accuracies of 95.06% and 96.14%, respec-tively,KNN showed comparatively lower recall (89.29%).Lo-gistic Regression highly improved performance with 99.43% accuracy and an F1 Score of 97.81% and has a balance between precision and recall.

TABLE III
COMPARATIVE PERFORMANCE OF MACHINE LEARNING MODELS (IN %)

Model	Accuracy	Precision	Recall	F1 Score
Linear Regression	92.75	92.34	80.31	80.21
Naive Bayes	95.06	91.74	93.54	94.24
KNN	96.14	94.95	89.29	92.67
Logistic Regression	99.43	99.24	93.25	97.81
XGBoost	99.59	98.94	97.56	98.29
SVM	99.69	99.66	98.26	99.29
Random Forest	99.72	99.73	99.58	99.51
Decision Tree	99.75	99.66	99.86	99.69

Across the different models, ensemble-based approaches achieved highest results. XGBoost, SVM, Random Forest and Decision Tree all exceeded 99% accuracy.Decision Tree achieved the highest accuracy (99.75%) and recall (99.86%).Random Forest showed high performance with an F1 Score of 99.51%.Overall,the findings show that tree-based and ensemble learning techniques have greater performance as compared to linear and probabilistic models for student performance prediction.

Table IV presents a comparative summary of the implemented machine learning algorithms. The models are regression-based, probabilistic,instance-based,tree-based, margin-based and ensemble approaches.

Based on the study results,ensemble and tree-based models show higher prediction performance.Decision Tree achieved the highest recall,Random Forest and XGBoost showed very high accuracy.SVM and Logistic Regression also performed well,whereas Linear Regression showed lower recall and F1 performance. Overall the advanced supervised and ensemble techniques were more effective for student performance pre-diction.

To analyze the classification performance of the models, confusion matrices are very useful.Figure 4 presents the con-fusion matrix of the Decision Tree model,the distribution of predicted and actual student performance are shown through it.

Figure 4 shows that most of the predictions lie along the diagonal,which indicates high accuracy classification across

TABLE IV
COMPARISON OF IMPLEMENTED MACHINE LEARNING ALGORITHMS

Algorithm	Type	Remarks Based on Study
Linear Regression	Regression	Baseline model,lower recall and F1 performance
Naive Bayes	Probabilistic	Fast and efficient,slightly lower precision
k-Nearest Neighbors (k-NN)	Instance	Balanced perfor-mance,sensitive to feature scaling
Logistic Regression	Linear Classifier	High accuracy with constant precision and recall
Decision Tree	Supervised/Tree Based	Achieved highest re-call
Random Forest	Ensemble	Very high accuracy and strong general-ization
Support Vector Machine (SVM)	Margin-based Classifier	Strong classification performance with high precision
XGBoost	Ensemble	High predictive per-formance

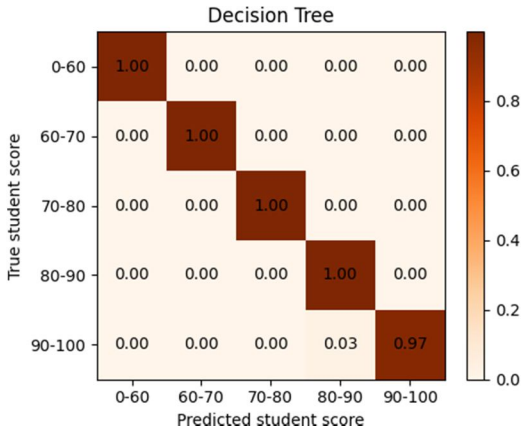


Fig. 4. Confusion matrix of Decision Tree

grade ranges. Only minimal misclassification is visible in the 90–100 category.

Confusion matrix of the Logistic Regression model is shown in Figure 5 compares model behaviour across grades.This visualization highlights the distribution of the predicted and actual student performance levels.

Figure 5 shows strong classification performance across most of the grade ranges, predictions are concentrated along the diagonal.
Figure 6 presents confusion matrix of Linear Regression model to analyze its classification capability across different score categories.As shown in Figure 6,the Linear Regression model has some noticeable misclassification, mainly in the higher score ranges (80–100). The dispersion away from the diagonal confirms that is has lower recall and F1 performance then other models

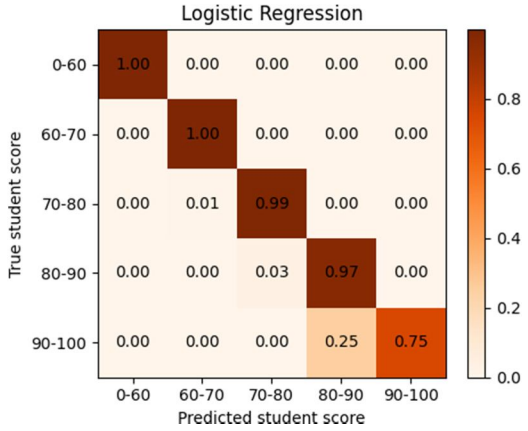


Fig. 5. Confusion matrix of Logistic Regression

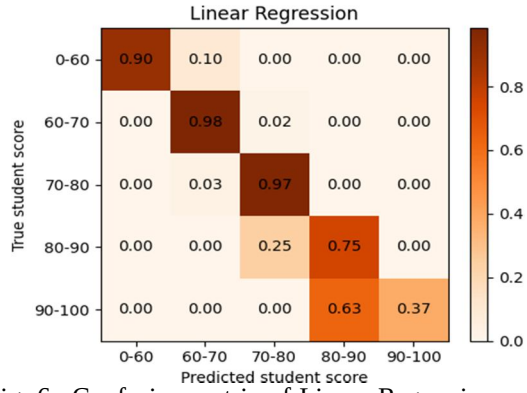


Fig. 6. Confusion matrix of Linear Regression

VII. CONCLUSION AND FUTURE WORK

The conducted study has explored various applications of machine learning techniques used for predicting academic performance of students using the educational data collected from academic,behavioral and demographic sources.The ex-perimental analysis shows that machine learning models are capable of identifying complex relationships between various factors related to students that are overlooked in traditional methods. Conventional assessment approaches that depend mainly on examination scores and manual judgment whereas the machine learning frameworks provide a data-oriented and systematic way to analyze student performance patterns.The results show that automated prediction models can give higher accuracy,consistency and scalability as they can be applied to large academic datasets.

Future research can address these limitations by extending the present frameworks in various manners.The integration of deep learning models like recurrent and attention-based networks,may increase the accuracy of prediction because they can model academic behavior sequentially.Real-time pre-diction systems can also be developed by including data

streams from LMS platforms, continuous monitoring of stu-dent progress throughout the semester is possible because of this.Expanding the dataset to include multi-institutional and longitudinal data in it would also increase the generalizability of the model.Moreover,future work should also focus on ethical considerations related to data privacy and transparency so that student data is used responsibly.By addressing these challenges,student performance prediction systems can be-come accurate and reliable tools in modern education system. One more improtant part in future research is including AI techniques into student performance prediction systems that can explain the reason of the results. While high prediction ac-curacy is desirable,educators and academic administrators also require transparency to understand how and why predictions are made. Explainable models can increase trust and ensure that automated predictions align with academic policies and ethical standards.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Finally, ethical considerations, fairness and responsible use of student information as well as privacy of educational data is important while developing predictive systems.Compliance with data protection regulations is must.Adopting to trans-parent data governance practices will be essential for sus-tainable deployment of machine learning solutions in educa-tion.Addressing these aspects will help to make student perfor-mance prediction systems reliable, ethical and also impactful to improve educational outcomes.

This study includes the following dataset to analyze the performance of machine learning algorithms.The dataset used is the *Student Performance Dataset*,which contains student academic records.The dataset includes attributes namely weekly self-study hours,attendance percentage,class participation,total score and final grade.These variables help to analyze the relationship between student learning behavior and academic outcomes.

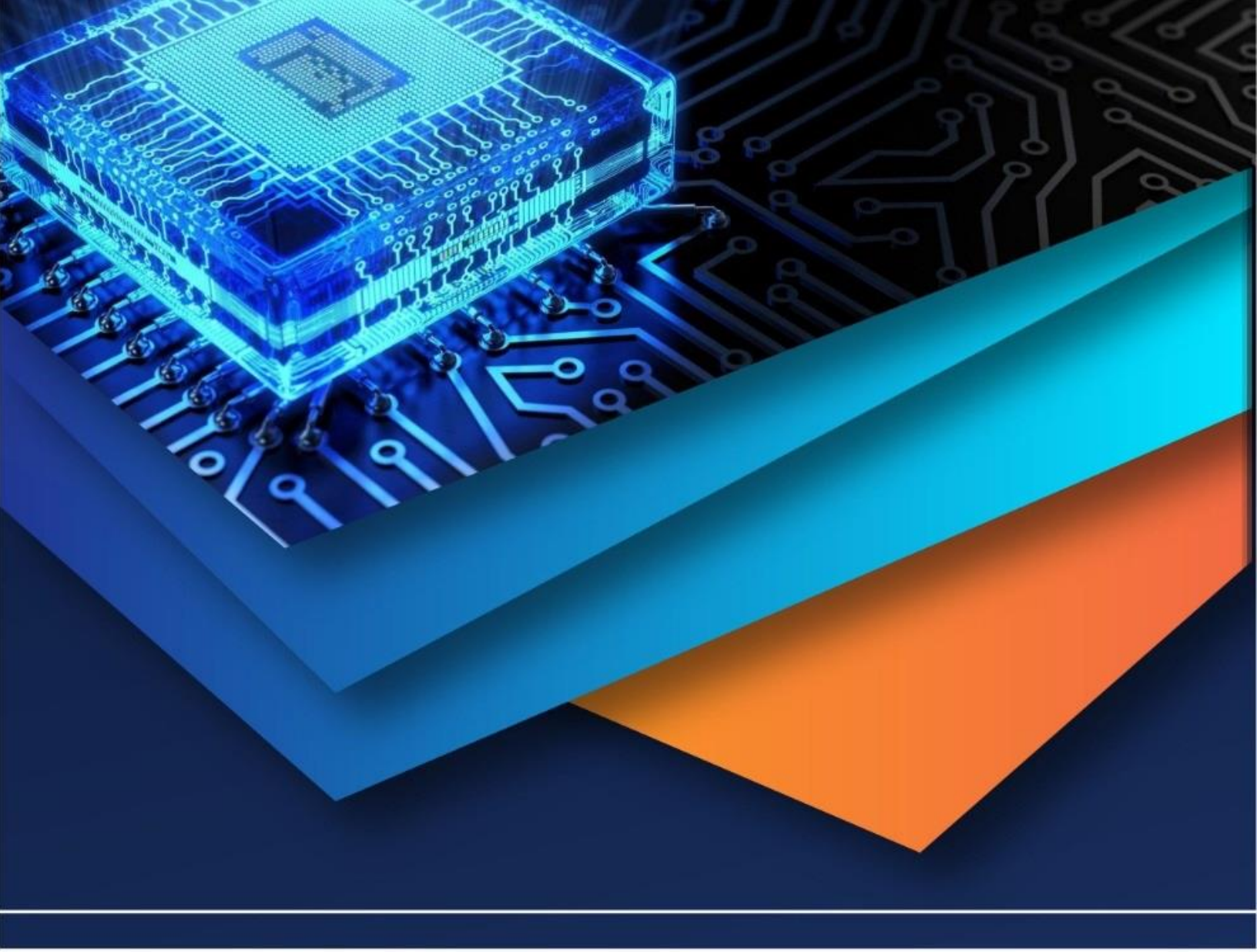
Dataset: Student Performance Dataset URL:https://www.kaggle.com/datasets/nabeelqureshitiii/ student-performance-dataset

REFERENCES

[1] W. Ahmed et al., "Learning-based academic performance prediction with explainability for enhanced decision-making in educational institutions," Discover Computing, 2025. https://link.springer.com
[2] M. N. Gul et al., "Data driven decisions in education using a comprehen-sive machine learning framework for student performance prediction," Discover Computing, 2025. https://link.springer.com
[3] N. F. A. Rahman et al., "Artificial intelligence in education: A systematic review of machine learning for predicting student performance," Journal of Advanced Research in Applied Sciences and Engineering Technology, 2025. https://akademiabaru.com
[4] C.-A. Buzducea et al., "Machine learning in education: Predict-ing student performance and guiding institutional decisions," 2024. https://www.researchgate.net
[5] E. Ahmed, "Student performance prediction using machine learning algorithms," 2024. https://www.researchgate.net
[6] H. Munir, B. Vogel, and A. Jacobsson, "Artificial intelligence and ma-chine learning approaches in digital education: A systematic revision," 2024. https://www.sciencedirect.com
[7] Y. N. S. Al Husaini and N. S. A. Shukor, "Factors affecting students' academic performance: A review," 2024. https://www.sciencedirect.com
[8] M. Yag˘cı, "Educational data mining: Prediction of students' aca-demic performance using machine learning algorithms," 2022. https://doi.org/10.1186/s40561-022-00192-z
[9] G. M. Rao and K. K. Kumar, "Students performance prediction in online courses using machine learning algorithms," 2021. https://uijrt.com
[10] B. S¸ ekeroglu et al., "Systematic literature review on machine learning and student performance prediction," 2021. https://doi.org/10.3390/app112210907
[11] B. Albreiki et al., "A systematic literature review of student performance prediction using machine learning techniques," 2021. https://doi.org/10.3390/educsci11090552
[12] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques," 2020. https://doi.org/10.3390/app11010237
[13] A. S. Hashim, "Student performance prediction model based on super-vised machine learning algorithms," 2020. https://doi.org/10.1088/1757-899X/928/3/032019
[14] A. A. Enughwure and M. E. Ogbise, "Application of machine learning methods to predict student performance," 2020. https://www.irjet.net
[15] H. Altabrawee et al., "Predicting students' performance using machine learning techniques," 2019. https://www.iasj.net
[16] B. S¸ ekeroglu et al., "Student performance prediction and classification using machine learning algorithms," 2019. https://dl.acm.org
[17] A. Namoun, "Educational analytics and early warning systems for student performance prediction," 2023. https://www.researchgate.net
[18] S. M. Ajibade et al., "Educational data mining: Enhancement of student performance model using ensemble methods," 2020. https://iopscience.iop.org
[19] L. Breiman, "Random forests," 2001. https://doi.org/10.1023/A:1010933404324
[20] C. Cortes and V. Vapnik, "Support-vector networks," 1995. https://doi.org/10.1007/BF00994018
[21] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," 2016. https://doi.org/10.1145/2939672.2939785
[22] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," 2001. https://doi.org/10.1214/aos/1013203451
[23] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," 2017. https://papers.nips.cc
[24] M. Ribeiro et al., "Why should I trust you? Explaining the predictions of any classifier," 2016. https://doi.org/10.1145/2939672.2939778
[25] I. Goodfellow et al., Deep Learning. https://www.deeplearningbook.org
[26] T. Hastie et al., The Elements of Statistical Learning. https://hastie.su.domains/ElemStatLearn/
[27] F. Pedregosa et al., "Scikit-learn: Machine learning in Python." https://jmlr.org/papers/v12/pedregosa11a.html
[28] A. Barocas et al., Fairness and Machine Learning. https://fairmlbook.org
[29] F. Trujillo, M. Pozo, and G. Suntaxi, "Artificial intelligence in education: A systematic literature review of machine learning approaches in student career prediction," Journal of Technology and Science Education, vol. 15, no. 1, pp. 162–185, 2025. https://doi.org/10.3926/jotse.2603
[30] S. M. F. D. Syed Mustapha, "Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods," Applied System Innovation, vol. 6, 2023. https://doi.org/10.3390/asi6030086
[31] P. Manjare, R. Shelke, and S. Kaware, "Comparative study of stu-dent academic outcomes and behavioral patterns through data-driven approaches," ITM Web of Conferences, vol. 81, 2026. https://www.itm-conferences.org/
[32] A. Gole, S. Singh, P. Kanherkar, P. R. Abhishek, and P. P. Wankhede, "Comparative analysis of machine learning algorithms: Random Forest, Naive Bayes classifier and KNN – A survey," International Journal for Research Publication & Seminar, vol. 13, no. 3, 2022. https://ijrps.org/
[33] Z. Sun et al., "Comparing machine learning models and statistical models for predicting heart failure events: A systematic review and meta-analysis," Frontiers in Cardiovascular Medicine, vol. 9, 2022. https://doi.org/10.3389/fcvm.2022.863842
[34] T. P. Pagano et al., "Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and mitigation methods," Big Data and Cognitive Computing, vol. 7, no. 15, 2023. https://doi.org/10.3390/bdcc7010015
[35] D. Markovics and M. J. Mayer, "Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction," Renewable and Sustainable Energy Reviews, vol. 161, 2022. https://doi.org/10.1016/j.rser.2022.112364
[36] T. Niu et al., "Decoding student cognitive abilities: A comparative study of explainable AI algorithms in educational data mining," Scientific Reports, vol. 15, 2025. https://www.nature.com/

[37] Q. Long et al., "Enhancing happiness and life satisfaction in university students: Analysis with a machine learning approach," BMC Psychology, vol. 14, 2026. https://bmcpsychology.biomedcentral.com/
[38] M. Jain and A. Srihari, "Comparison of machine learning algorithms in intrusion detection systems: A review using binary logistic regression," International Journal of Computer Science and Mobile Computing, vol. 13, no. 10, pp. 45–53, 2024. https://ijcsmc.com/
[39] N. Al-Shanableh et al., "Forecasting students' academic performance in educational data using machine learning techniques," International Journal of Information and Communication Technology Education, vol. 22, no. 1, 2026. https://www.igi-global.com/
[40] J. K. Rogers, T. C. Mercado, and R. Cheng, "Predicting student performance using Moodle data and machine learning with feature im-portance," Indonesian Journal of Electrical Engineering and Computer Science, vol. 37, no. 1, pp. 223–231, 2025. https://ijeecs.iaescore.com/
[41] Y. Chen et al., "Machine learning-driven student performance prediction for enhancing tiered instruction," arXiv preprint, 2025. https://arxiv.org/
[42] S. Wang et al., "Artificial intelligence in education: A systematic literature review," Expert Systems with Applications, vol. 252, 2024. https://doi.org/10.1016/j.eswa.2024.124167
[43] M. A. Cardona, R. J. Rodr´ıguez, and K. Ishmael, "Artificial intelligence and the future of teaching and learning: Insights and recommendations," U.S. Department of Education, 2023. https://www.ed.gov/
[44] S. Somvanshi et al., "From tiny machine learning to tiny deep learning: A survey," ACM Computing Surveys, 2026. https://dl.acm.org/
[45] M. Mart´ınez-Comesan˜a et al., "Impact of artificial intelligence on assessment methods in primary and secondary education: System-atic literature review," Revista de Psicodida´ctica, vol. 28, 2023. https://doi.org/10.1016/j.psicoe.2022.11.001
[46] J. Wang and Y. Yu, "Machine learning approach to student perfor-mance prediction of online learning," PLOS ONE, vol. 20, 2025. https://doi.org/10.1371/journal.pone.0300000
[47] G. Wei, G.-S. Han, and X. Lang, "Fire risk assessment using machine learning techniques: A case study of Jinan City, China," Scientific Reports, 2026. https://www.nature.com/
[48] S. Albugami, A. Wali, and H. Almagrabi, "Predicting student dropout in Saudi universities using machine learning and explainable AI," PeerJ Computer Science, 2026. https://peerj.com/
[49] A. M. Vieriu and G. Petrea, "The impact of artificial intelligence (AI) on students' academic development," Education Sciences, vol. 15, no. 3, 2025. https://doi.org/10.3390/educsci15030343
[50] A. I. AbuEid et al., "Artificial intelligence in education predicting college plans of high school students," Journal of Intelligent System and Applied Data Science, vol. 2, no. 1, 2024. https://jisads.org

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)